LBID-2512

National Energy Research Scientific Computing (NERSC) Center: From Here to Petaflop/s

A Proposal to the DOE Office of Science from Lawrence Berkeley National Laboratory

June 30, 2004

Horst Simon, William Kramer, William Saphir, John Shalf, David Bailey, Leonid Oliker

PI: Horst D. Simon, Associate Laboratory Director for Computing Sciences
Mail Stop 50B-4230, Lawrence Berkeley National Laboratory, Berkeley, CA 94720
(510) 486-7377, fax (510) 486-4300, hdsimon@lbl.gov

This work was supported by the Director, Office of Science, Office of Advanced Scientific Computing Research of the U.S. Department of Energy under Contract No. DE-AC 03-76SF00098.

Table of Contents

Executive Summary

Lawrence Berkeley National Laboratory (Berkeley Lab) proposes to expand the computational capability of the National Energy Research Scientific Computing (NERSC) Center to provide leadership-class computing resources and to establish a new partnership between the American computer industry and a national collaboration of laboratories, universities, and computing facilities. NERSC will provide leadership-class scientific computing capability to scientists and engineers nationwide, independent of their institutional affiliation or source of funding. This partnership will bring into existence a new class of computational capability in the United States that is optimal for science and will create a sustainable path towards petaflop/s performance. NERSC's proposal is consistent with the goals of the "Federal Plan for High-End Computing," the report of the High-End Computing Revitalization Task Force (HECRTF) [1], and will lead to a petaflop/s capability for DOE applications by the end of the decade.

NERSC will change the way that high performance computers are designed and deployed through a new type of development partnership with the computer vendor based on the concept of *science-driven computer architecture*. We propose a partnership with IBM, the leading U.S. vendor of high performance computing systems for science, to design a science-driven computer architecture with an aggregate sustained performance of 100+ Tflop/s on a broad spectrum of applications of national importance in FY 2008. This will be achieved through the phased development and installation of two leadership-class systems (NERSC-5L and NERSC-6L) and other systems to support them. These systems will provide a factor of 100 improvement over current performance levels, and thus provide the leadership called for in the HECRTF report [1, pg. 29]. The proposal culminates with a petaflop/s computing system for production-quality science available in 2010. The systems will be designed with assistance from computational scientists, with prototype pre-commercial configuration and testing under way at IBM. This architectural approach achieves the highest sustained performance across a broad range of key scientific applications for the lowest cost. It provides the best national investment for scientific productivity, demonstrates continued U.S. leadership in computational science, and forges a path to petaflop/s computing [see 1, pg. 14].

Applications scientists have been frustrated by a trend of stagnating application performance, despite dramatic increases in claimed peak performance of high-performance computing systems. This trend, often referred to as the "divergence problem" [see 1, pg. 13], will be directly countered by our strategy. Our strategy reverses that trend by engaging applications scientists well before an architecture is available for commercialization. The partnership with IBM is based on a collaborative approach to designing computer architecture that will enable heretofore unrealized achievements in computer-capability-limited fields, including nanoscience, combustion modeling, fusion, climate modeling, and astrophysics. The unprecedented level of computational capability that will be made available to researchers in these fields will result in scientific breakthroughs on issues of national importance.

NERSC has a demonstrated track record in acquiring and fielding high-end systems that meet user requirements and lead the country in unclassified scientific computing. Leveraging the expertise at NERSC will significantly reduce operational costs of leadership computing, while ensuring a timely installation and operation of the new systems.

The establishment of leadership-class computing at NERSC will result in a scientific computing capability that balances well the investment of the country in leadership-class machines, while providing much-needed resources to computational scientists working on DOE mission-critical applications. It complements hardware investments made elsewhere in DOE by providing an alternative architecture route, thus mitigating risk and increasing architectural diversity. The proposed class of computer designs will not only revolutionize the power of supercomputing for science, but it also will affect scientific computing at all scales. This proposal represents our vision for achieving outstanding computational science by providing for the continued development of science-driven computer architecture.

1. INTRODUCTION: A SUSTAINABLE PATH TO LEADERSHIP COMPUTING

The National Energy Research Scientific Computing (NERSC) Center is the premier open and unclassified computing facility for the Office of Science. Operating the NERSC Center has enabled Lawrence Berkeley National Laboratory (Berkeley Lab) to acquire unsurpassed expertise in operating large computational and storage systems, integrating them into high-speed networks, and providing comprehensive scientific support that enables researchers to make the most productive use of these resources. NERSC supports more than 2,400 users nationally and internationally. Over 50% of the users are from universities. NERSC's success is measured by the scientific productivity of its users.[1] Its staff and management are adept at balancing and satisfying the diverse needs of researchers within the constraints imposed by programmatic missions, goals, and requirements. NERSC is known worldwide for the quality of its computing services.

With the exception of the science-driven architecture activity described in Section 3.5, the systems and services discussed in this proposal will be supported by the existing NERSC staff resources — approximately 65 technical FTE. This staff is sufficient to continue all the services at NERSC and extend them to the leadership-class systems. The details of the services NERSC staff provide are in Appendix F of [2].

NERSC's user services and scientific support are highly regarded throughout the HEC community. NERSC provides a 24-by-7 help desk where it is possible for a user to talk to NERSC staff. They also produce leading-edge training and state-of-the-art documentation for NERSC systems. The NERSC support staff provide specialized services ranging from supporting unique software needs, to special processing and scheduling, to long-term collaborative interaction in order to build and optimize science codes. The support staff are highly trained (most are Ph.D.s), not just in computational methods but also in scientific disciplines.

NERSC has a sophisticated account management and allocation management system that has automated many routine tasks. It gives instantaneous access to usage data, not just for users and PIs but for DOE program managers.

NERSC proactively engages with the user community through the NERSC User Group, which meets monthly by phone and semi-annually face to face. NERSC also measures a number of quality metrics ranging from typical reliability/availability/serviceability measures to job throughput, system efficiency, and responsiveness solving problems reported by users. The most telling evidence is the annual NERSC User Survey that assesses user satisfaction.[2] One user characterized the quality of NERSC as follows:

> *"NERSC simply is the best-run centralized computer center on the planet. I have interacted with many central computer centers and none are as responsive, have people with the technical knowledge available to answer questions, and have the system/software as well configured as does NERSC." —2003 NERSC User Survey Respondent*

This proposal presents a plan that will maintain and strengthen U.S. leadership in high performance computing, initiate a new wave of scientific discovery, and enable the solution of problems of national importance. Berkeley Lab proposes expanding NERSC to provide leadership-class computing resources and to establish a partnership between the American computer industry and a national collaboration of laboratories, universities, and computing facilities. In this proposal we are guided by the following analysis:

[1] For 96 pages of citations of publications resulting from computations at NERSC in 2003, see http://www.nersc.gov/news/reports/ERCAPpubs03.php.
[2] For the latest survey results, see http://www.nersc.gov/news/survey/2003/.

1. This U.S. Department of Energy (DOE) investment must lead to widely deployable new technology for high-end scientific computing. If it leads merely to a series of experiments or the purchase of a single machine, it will not have a lasting impact.
2. The technology we need will not spontaneously appear on the market. By taking a passive approach that relies on existing vendor offerings, the high performance computing community has ceded leadership to other players whose requirements are increasingly incompatible with the needs of high-end computing.
3. Several national panels have concluded that the rules of engagement between the scientific community and the American computer industry must be revised. Scientific applications must directly influence machine design in a repeating cycle: (a) scientific applications input to designers, (b) computer design with increased performance, (c) deployment and delivery of new systems to the scientific community, (d) repeat.
4. Successfully changing the rules of engagement requires a partnership with the American computer company with the most resources, the best track record of research and development, and proven success in delivering in high performance computing. To justify the necessary commitment from the company, we will form a national collaboration of laboratories, computing facilities, universities, and researchers equally committed to changing the future of the computing capability available to the scientific community.
5. Berkeley Lab and our partners have evaluated a representative array of scientific applications to establish precisely their algorithmic characteristics. From those algorithms we have derived a clear understanding of the limitations of current high-end systems of all designs, from clusters to vector computers.
6. Over the past two years, the Blue Planet partnership led by Berkeley Lab has worked closely with IBM to design a machine that better meets the needs of scientific applications. The goals and methodology of this partnership were validated by the successful design and implementation of the $100M ASCI Purple system at Lawrence Livermore National Laboratory (LLNL), based on the Blue Planet design.
7. We propose to continue and expand the Blue Planet process, bringing in appropriate partners to guide the process. IBM has committed to participate, delivering a leadership-class computing system based on an extension of the Blue Planet work, with possible evolution into a hybrid with the IBM Blue Gene designs currently under research and development.
8. Berkeley Lab, through its effective management of NERSC, has earned the reputation for delivering the best high-end computing to the national scientific community. Leveraging these resources and experience will produce the greatest return on DOE's investment and provide the greatest opportunity for a successful scientific program.

NERSC has already started on the path to deploying the optimal science-driven architecture and building the national collaboration necessary for the successful realization of this vision.

2. SCIENTIFIC APPLICATIONS AND UNDERLYING ALGORITHMS DRIVE ARCHITECTURAL DESIGN

The central goal of this proposal is to deliver new scientific results on computations of a scale that greatly exceeds what is possible on current systems, with sustained aggregate performance rates of 100+ Tflop/s in 2008 on applications of scientific and national importance. To that end, we have identified the following application classes as being ripe for breakthrough science using very high-end computing, and relevant to some of the most important national objectives: nanoscience, combustion modeling, fusion energy simulations, climate modeling, and astrophysics. More application classes are likely to use the facility as well. Table 1 summarizes the goals, computational methods, and example applications of each science area. These goals are DOE mission-relevant goals and consistent with the potential breakthroughs listed in Table 1-A and B of the HECRTF report [1].

Table 1
Science Breakthroughs Enabled by Leadership Computing Capability

Science Areas	Goals	Computational Methods	Examples of Breakthrough Applications
Nanoscience	Simulate the synthesis and predict the properties of multi-component nanosystems	Quantum molecular dynamics Quantum Monte Carlo Iterative eigensolvers Dense linear algebra Parallel 3D FFTs	Simulate nanostructures with hundreds to thousands of atoms, as well as transport and optical properties and other parameters
Combustion Modeling	Predict combustion processes to provide efficient, clean and sustainable energy	Explicit finite difference Implicit finite difference Zero-dimensional physics Adaptive mesh refinement Lagrangian particle methods	Simulate laboratory-scale flames with high-fidelity representations of governing physical processes
Fusion Energy	Understand high-energy density plasmas and develop an integrated simulation of a fusion reactor	Multi-physics, multi-scale Particle methods Regular & irregular access Nonlinear solvers Adaptive mesh refinement	Simulate the ITER reactor
Climate Modeling	Accurately detect and attribute climate change, predict future climate, and engineer mitigation strategies	Finite difference methods FFTs Regular & irregular access Simulation ensembles	Perform a full ocean/atmosphere climate model with 0.125 degree spacing, with an ensemble of 8–10 runs
Astrophysics	Determine through simulation and analysis of observational data the origin, evolution, and fate of the universe; the nature of matter and energy; galaxy and stellar evolution	Multi-physics, multi-scale Dense linear algebra Parallel 3D FFTs Spherical transforms Particle methods Adaptive mesh refinement	Simulate the explosion of a supernova with a full 3D model

The most effective approach to designing a computer architecture that can meet these scientific needs is to analyze the underlying algorithms of these applications, and then, working in partnership with vendors, design a system targeted to these algorithms.

From this list of important scientific applications and underlying algorithms, several themes can be derived that drive the choice of a large-scale scientific computer system: (1) multi-physics, multi-scale calculations; (2) limited concurrency, requiring strong single-CPU performance; (3) reliance on key library routines such as ScaLAPACK and FFTs; (4) the use of particle methods, with couplings to grid-based methods that lead to large-scale interaction of two regular, but unaligned, data structures; (5)

widespread usage of finite difference computations, requiring good performance on fairly regular accesses in multiple dimensions and high main memory bandwidth; (6) an increasing usage of sparse, unstructured, and adaptive mesh (AMR) methods, which entail some irregular control sequences that do not perform well on vector systems; and (7) ubiquitous data parallelism providing the opportunity for fine-grained operation concurrency; (8) irregular control flow inhibiting fine-grained symmetric operation concurrency. Table 2 presents a qualitative summary of this information:

Table 2
Algorithm Requirements

Science Areas	Multi-physics & multi-scale	Dense linear algebra	FFTs	Particle methods	AMR	Data parallelism	Irregular control flow
Nanoscience	X	X	X	X		X	X
Combustion	X			X	X	X	X
Fusion	X	X		X	X	X	X
Climate	X		X		X	X	X
Astrophysics	X	X	X	X	X	X	X

The characteristics summarized here point to the need for a flexible system — one that can perform well both on random memory access calculations as well as regular memory access problems and that combines strong single-node performance (to minimize the required concurrency in the application) and a powerful system-scale network.

Of the two principal classes of high performance systems in widespread usage — superscalar systems and vector systems — each has a different set of advantages and disadvantages for these applications. Superscalar, cache-memory-based systems tend to do well on problems with spatial and temporal data regularity. These systems also do relatively well on irregularly structured algorithms and codes with heavy usage of conditional branching in inner loops. However, many cache-based systems feature low or over-subscribed main memory bandwidth, since they are not primarily designed for scientific computation. Thus, codes with low computational intensity typically do not perform well on these architectures.

Vector systems exploit regularities in the computational structure to expedite uniform operations on dependence-free data. Some scientific codes are characterized by predictable fine-grained data-parallelism and thus allow vectorization. However, vector systems tend to do poorly on codes with irregularly structured computations. These codes are characterized by irregular control flow, intensive scalar operations, and significant conditional branching — operations that inhibit vectorization. Performance on vector architectures degrades significantly even when a small fraction of the work is non-vectorizable, as described by Amdahl's Law. This is particularly true for newly emerging multi-method, multi-physics codes that can only leverage vectorization for a subset of the numerical components.

These considerations suggest that an architecture that combines the best features of high-end superscalar and vector systems would be best suited for the workload that we project for future high-end computing of national importance. To that end, we will describe in the following sections a system that is being developed by IBM, in collaboration with NERSC, that targets this broad range of scientific computing.

3. A SCIENCE-DRIVEN SYSTEM ARCHITECTURE

Applications scientists have been frustrated by a trend of stagnating application performance despite dramatic increases in claimed peak performance of high-performance computing (HPC) systems. This trend has been widely attributed to the use of commodity components whose architectural designs are unbalanced and inefficient for large-scale scientific computations [1]. It was assumed that the ever-increasing gap between theoretical peak and sustained performance was unavoidable. However, recent results from the Earth Simulator (ES) in Japan clearly demonstrate that a close collaboration with a vendor to develop a science-driven architectural solution can produce a system that achieves a significant fraction of peak performance for critical scientific applications. The key to the ES success was the long-term collaborative development strategy between the scientists of JAMSTEC (Japan Marine Science and Technology Center) and NEC Corporation.

Realizing that effective large-scale system performance cannot be achieved without a sustained focus on application-specific architectural development, NERSC and IBM have led a collaboration since 2002 that involves extensive interactions between domain scientists, mathematicians, computer experts, as well as leading members of IBM's R&D and product development teams. The goal of this effort is to change IBM's architectural roadmap to improve system balance and to add key architectural features that address the requirements of demanding leadership-class applications — ultimately leading to a sustained Pflop/s system for scientific discovery. The first product of this multi-year effort has been a redesigned Power5-based HPC system known as Blue Planet [3] and a set of architectural extensions referred to as ViVA (Virtual Vector Architecture). This collaboration has already had a dramatic impact on the architectural design of the ASCI Purple system [4], and has resulted directly in the strong NERSC leadership-class systems (NERSC-5L and NERSC-6L) presented in this proposal.

Blue Planet design is incorporated into the new generation of IBM Power microprocessors that are the building blocks of the NERSC-5L and NERSC-6L configurations. These processors break the memory bandwidth bottleneck, reversing the recent trend towards architectures poorly balanced for scientific computations. The Blue Planet design improved the original power roadmap in several key respects: dramatically improved memory bandwidth; 70% reduction in memory latency; eight-fold improvement in interconnect bandwidth per processor; and ViVA Virtual Processor extensions, which allow all eight processors within a node to be effectively utilized as a single virtual processor.

The approach described in this proposal — a multi-stage deployment of a leadership-class system, NERSC-5L and NERSC-6L — is a continuation of the work that started in the Blue Planet initiative. We propose what will probably be the first petaflop/s computing system for production-quality science in 2010 — NERSC-7P. Finally, in order to achieve these aggressive goals, we include a persistent science-driven architecture process that will assure the success of ViVA-2 in NERSC-6L and prepare for the 2010 petaflop/s system. We will expand upon this successful collaborative effort, starting with the baseline configurations discussed below. The purpose of this collaborative approach is not just to produce the most effective scientific computing platform in the NERSC-6L timeframe, but also to begin moving on a longer-term roadmap towards successful petaflop/s computing.

3.1 Building on the Blue Planet Collaboration: Addressing the Memory Bandwidth Bottleneck

We propose to continue and expand the Blue Planet process to develop further improvements to the NERSC-6L system and beyond. Note that past efforts of LBNL and LLNL have greatly influenced NERSC-5L already. This continued collaboration will lead to a set of enhancements known as ViVA-2. We will hold workshops to define the issues of leadership-class computing, and semi-annual meetings with the NERSC users and advisors to review progress, create ideas, and refine the design decisions. These meetings will integrate application scientists, system designers, HPC performance experts, and computer scientists. This community approach of directly engaging vendors in the collaborative process

of designing leadership HPC systems was laid out by the High End Computing Revitalization Task Force (HECRTF) Workshop [5], the Federal Plan for High-End Computing [1], and the DOE SCaLeS Workshop [6], and has been demonstrated successfully by the Earth Simulator, the initial Blue Planet effort, and the Red Storm effort [7].

There is an opportunity to incorporate the ViVA-2 scientific enhancement technology into future Power processor design. During FY04 and FY05, IBM and NERSC, along with other partners, will evaluate various enhancements to the NERSC-6L processor, node, and interconnect design, including assisted processing capabilities and their impact on the associated components (e.g., compilers, libraries, tools, etc.). Then NERSC will advise IBM on how to incorporate the resulting technology into NERSC-6L and subsequent systems, to maximize its impact on scientific discovery. Thus, the NERSC-6L system described in this document should be considered a minimum base from which improvements will evolve.

IBM's willingness to work with NERSC[3] to develop modifications to its hardware that further enhance performance of scientific applications clearly demonstrates their commitment to scientific computing and the importance of IBM's partnership with the computational science community. IBM is the only company that both demonstrates a clear commitment to make such deep changes to their design and offers the immense resources required to meet those commitments.

ViVA Design Targets

ViVA and ViVA-2 are specialized enhancements to the Power architecture designed to significantly improve sustained performance on a wide range of scientific applications. ViVA is a compiler-supported programming model that combines processors to form more powerful virtual processors by making use of fast barrier synchronization technology available in Power5 and Power6 processors. ViVA will be available on both the NERSC-5L and NERSC-6L systems.

ViVA-2 is envisioned as a set of extensions to the Power6 architecture that will accelerate scientific applications by supporting deeper pipelining of memory requests in order to hide memory latencies. These extensions will improve the efficiency of memory accesses on both vectorizable and non-vectorizable codes. ViVA-2 is superior to strictly vector designs because it offers the flexibility of achieving high performance on non-vectorizable algorithms using state-of-the-art superscalar technology, while efficiently processing data-parallel code segments that are amenable to vectorization. These enhancements address a variety of scalar memory performance degradations often attributed to irregularities in the data-access patterns. Examples include ineffective hardware prefetching, load/store instruction issue-rate limitations, and wasted bandwidth due to partially used cache lines.

The performance commitments stated in this proposal for NERSC-5L, NCS-L, and NERSC-6L do not depend on the successful implementation of ViVA and ViVA-2, and will be achieved in any case.

3.2 NERSC-5/6L and NERSC-7P: Leadership Computing Systems

Our goal is to build an architecture balanced for leadership-class science requirements as described above in Section 2, which presents the computational science applications that will be of critical importance to U.S. scientific leadership in 2008 and beyond and are able to take advantage of an ultra-scale computing system.

The key science requirements for leadership-class computing can be distilled into three main system features: processor performance, interconnect performance, and software. Processors should have excellent sustained single-node performance across the spectrum of applications. The interconnect should

[3] See letter from IBM in Appendix D of the proposal "National Facility for Advanced Computational Science: A Sustainable Path to Scientific Discovery" [2].

provide high per-link performance (both latency and bandwidth) as well as high bisection bandwidth. Effective system utilization requires proven system software scalability and optimized numerical libraries.

The goal of NERSC is to enable new science discoveries from a diverse population of computational scientists in a wide range of disciplines. Implicit in this is a requirement for real working systems. Our plans take into account both credibility and risk in vendor roadmaps. NERSC recently released a request for information (RFI) to the entire high performance computing and storage industry. The RFI went to over 40 high-end computing vendors.

From an analysis of the responses, we concluded that there are only two U.S. vendors with a credible roadmap to provide leadership-class computing capability in the 2008 time frame: IBM (Power6) and Cray (vector systems). Other vendors have competitive offerings at smaller scales, but not for the largest system scales. Additionally, software for cluster architectures is not sufficiently robust at this time to effectively manage a leadership-scale system.

After analyzing the latest system architecture and pricing information from IBM and Cray, we concluded that the IBM solution will ultimately be the best way to meet the need for a general-purpose leadership-class system between now and 2008 that satisfies the requirements of the NERSC applications base. With plans for Cray-based leadership systems with vector architectures at Oak Ridge National Laboratory and an IBM solution at NERSC, the Office of Science will be engaging the two major U.S. high-end computing vendors. This dual vendor, dual site strategy mitigates risks and provides increased architectural diversity for the SC community.

We have had access to an early Power5 system to run benchmarks and validate our assumptions about the ability of this processor to sustain a relatively high percentage of peak performance. Our tests confirmed that the Power5 will sustain high performance. Details of the technical rationale and benchmark results are discussed in Appendix A.

We propose a three-phased approach — NERSC-5L, NCS-L, and NERSC-6L — to achieve a leadership-class system in 2007 with an aggregate sustained performance of 100+ Tflop/s. NERSC-5L will be installed in June 2005, NCS-L in March 2006, and NERSC-6L in March 2008. NERSC-7P, the first petaflop/s system for production-quality science, will arrive in 2010.

NERSC-5L

NERSC-5L is a Power5 system with eight single-core CPUs per node and 1,024 nodes (8,192 CPUs). The CPUs run at 1.9 GHz (7.6 Gflop/s peak). The system will feature more than 62 Tflop/s peak and 12.5 Tflop/s average sustained performance. The system will have 32 terabytes (TB) of main memory and 640 TB of disk. The interconnect will be IBM's high performance Federation switch. It is expected that average application performance will be at least 16% of peak, with several key applications well above that range. Key innovations in the Power5 architecture that allow it to obtain a much higher percentage of peak performance than its predecessors, such as the Power4, include:

- **High-memory bandwidth per processor,** including a memory architecture that achieves 2.1 bytes/flop, comparable to vector architectures.
- **"Single core" node design.** IBM's original roadmap called for two processor cores on a single chip to share the same memory system. Going to a single core design effectively doubles the memory bandwidth per processor.
- **Small node design.** With eight-processor nodes, it is possible to put the processors closer to memory, reducing memory latency. Furthermore, by reducing the number of processors per node, effective network bandwidth per processor exceeds IBM's original 32- or 64-way SMP roadmap.
- **ViVA Virtual Processing** that allows the eight processors in a node to be treated as a single processor with peak performance of 61 Gflop/s. Codes that benefit from Cray X1 multistreaming,

for example, will directly benefit from ViVA capabilities. (See Appendix A of [2] for more details.)

The NERSC-5L network will be based on IBM's "Federation" interconnect. Two "planes" of this network will provide 8 GB/s of bidirectional bandwidth per node, or 1 GB/s per processor. Federation topology is a modified fat-tree that provides full-bisection bandwidth. Unlike systems that employ mesh and torus networks, the fat-tree network allows any processor to communicate with any other processor in the system free of bandwidth contention. This offers the most flexibility and the highest performance of any comparable system, resulting in a gross bisection bandwidth of 4 TB/s for NERSC-5L.

The global file system on the NERSC-L systems will be IBM's General Parallel File System (GPFS), a mature parallel file system that provides excellent performance and functionality. GPFS is the only parallel file system that has been demonstrated to support a diverse scientific parallel workload at the scale of multiple leadership-class systems.

A robust software environment is a critical component of a leadership architecture. System software on the NERSC-L systems will be an improved version of IBM's current cluster system software, which powers 45% of the Top 500 computers in the world and is the result of thousands of person years of effort. IBM's cluster software has proven its robustness and reliability at NERSC by consistently enabling utilization of 90 to 95% of available computational resources.

NERSC-L systems will have optimized mathematical and scientific libraries, including ESSL, MASS, and FFTW. Many codes poised to run at this scale depend on the availability of such libraries to extract maximum performance from the architecture.

NERSC-6L

NERSC-6L is a Power6 system with eight single-core CPUs per node, running at 5.0 GHz (20 Gflop/s). It has 2,048 nodes, for a total of 16,384 processors, with a total of 131 TB of main memory and 3.3 petabytes (PB) of shared global disk. The system will feature 327 Tflop/s peak and 79 Tflop/s average sustained performance. The Power6-based NERSC-6L system will have an impressive memory performance of 3.6 bytes/flop (72 GB/s per processor), allowing increased sustained performance across a broad spectrum of leading scientific applications.

The NERSC-6L network will be based on InfiniBand (IB) technology, an emerging industry standard that can be scaled to very high data rates. The NERSC-6L network will either be a single plane of IB 12x quad data rate, or two planes of IB 12x dual data rate, in either case providing approximately 24 GB/s of bi-directional bandwidth per node, or 3 GB/s per processor. The aggregate bandwidth of the full bisection network achieves 25 TB/s, allowing for efficient execution of large-scale applications with global communication requirements.

NERSC-6L will have the same basic file system and software as NERSC-5L, with improvements and larger scale.

NERSC-6L Refinements and Beyond

The ViVA-2 extensions being studied for NERSC-6L are intended to benefit scientific codes that are characterized by the kind of predictable data parallelism that is typically associated with vector processing. Since the superscalar core performs all computations on operands fetched by ViVA-2, its advantages are available even for non-vectorizable algorithms. NERSC will investigate design tradeoffs in collaboration with IBM and define the final ViVA-2 architecture.

Additionally, IBM is developing custom hardware accelerators in the network adaptors (HCA) to efficiently support collective operations and global barrier synchronizations in the NERSC-6L timeframe, specifically for leadership-class architectures. Specialized hardware support for global operations would result in significant reduction in latency overhead. These interconnect enhancements allow the NERSC-

6L to efficiently handle state-of-the-art scientific applications with fast global synchronization requirements, in a scalable fashion.

Based on the expertise gained from NERSC-6L system design, and the extensive application knowledge represented by the application partners, we will leverage the collaborative effort to assess the most effective and timely system options for a sustained Pflop/s system. IBM currently has the most diverse HPC research portfolio of any company in the world, including: BlueGene/L, DARPA HPCS PERCS, "cell" (Playstation-3) microprocessor technology, and Osmosis optical interconnect. The current roadmap, which is from IBM's RFI response, is described in Appendix B of [2] and depicted in Figure 1. NERSC will be involved early in this process in order to drive IBM and the community to an effective Pflop/s design for state-of-the-art scientific applications.

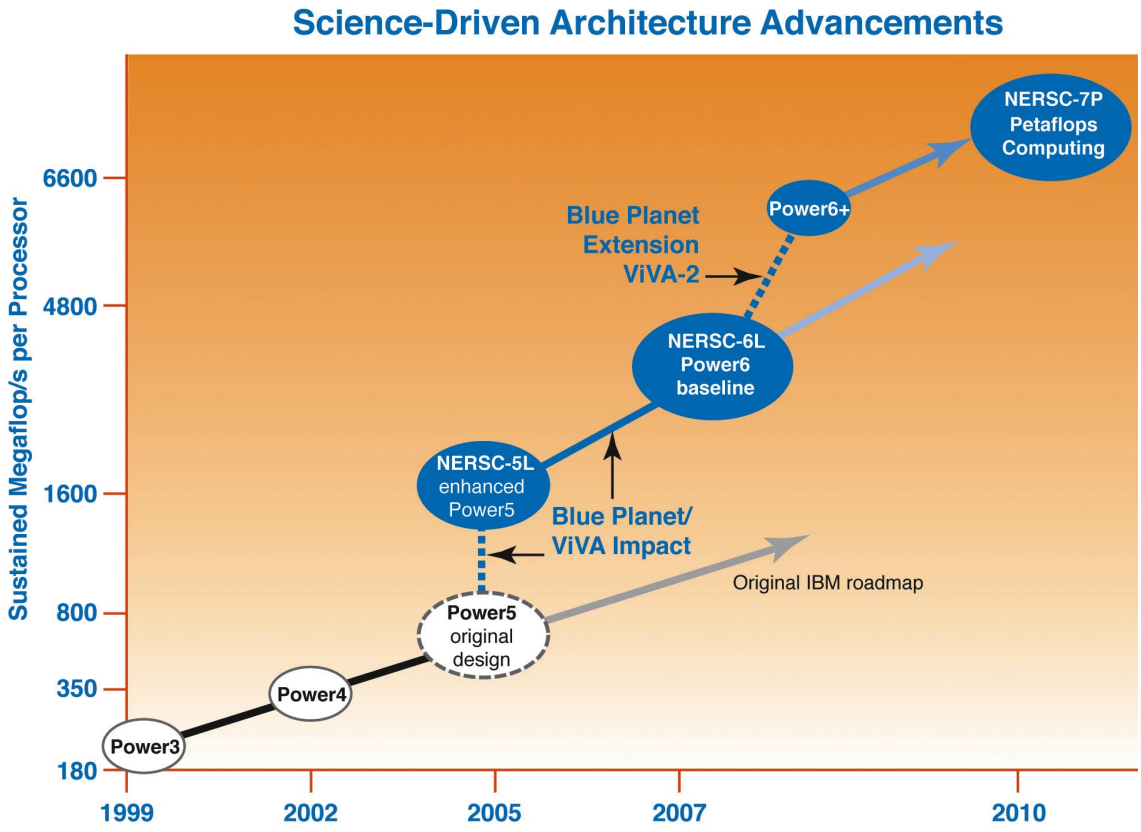

Figure 1. Science-driven architecture advancements.

3.3 NERSC Capacity System (NCS-L): A Leadership Commodity Cluster

The requirements from the NERSC science users make it clear that there is a significant need to have computational resources for applications that do not need the unique high-end resources of the largest system all the time. Even within leadership science projects, there is often the need to do smaller-scale processing for preparing data and analysis, post-processing results, exploratory ("back of the envelope") calculations, and evolving new algorithmic approaches that are much more efficient but may not scale fully at initial implementation (e.g., AMR). Further, there are projects that need a large volume of computational time for many jobs, but individual jobs do not require leadership resources — for example, the ensemble runs done by the climate community to determine the statistical properties of simulation

results. Other science areas (e.g., fusion) are now moving to this type of computing that augments the leadership-scale problems they are working on.

DOE recognizes this need and directed NERSC to implement a new computing system in FY 2004 as a new capability to address these requirements. This was a major topic at the recent NERSC User Group meeting, held June 25, which defined the issues to be addressed in the next "Greenbook" of user requirements. One of the main points from the scientists is that any such resource has to be proportionally scaled (1/5 to 4/5) to the largest resource. Hence, we propose adding NCS-L, a resource that is roughly scaled to NERSC-5L and NERSC-6L. The purpose of the NCS-L system is to satisfy the needs of a broad user base by providing a system configuration that suits particular subsets of the general NERSC workload, making more time available for the most demanding applications on the NERSC-5L and NERSC-6L systems.

NCS-L will be a parallel cluster based on commodity CPUs and commodity interconnect. Because it is commodity based, NCS-L is expected to provide a very cost-effective solution for the targeted classes of use. The system will be sized to take up the remaining space in Berkeley Lab's Oakland Scientific Facility (OSF) after NERSC-5L is installed. NERSC will evaluate several commodity configurations including the PowerPC (as in the Apple G5 series of processors), Intel's Irwindale (IA-32 architecture with 64-bit extensions), and AMD's Opteron CPUs. Interconnects considered include Mryinet 10G and Infiniband 4X. The IBM BladeCenter 2 system with Intel's Irwindale CPUs is currently considered the best solution, being the most cost effective and having a modest footprint.

In February 2004, Intel announced a major new thrust to its IA-32 processor architecture. The first implementation, code-named Nocona, adds Intel's Extended Memory 64 Technology, or EM64T. The Irwindale processor proposed here is the next version of Nocona, with faster clock speed and double the L2 cache size. Platforms with EM64T are binary compatible with 32-bit applications and can run 64-bit applications. EMT processors have twice as many registers as the IA-32 processors, and the processors are 64 and 128 bits in width. At 16 KB, Nocona's L1 cache is two times larger; it also includes 33% larger trace cache, and a new, more efficient branch prediction mechanism. Other enhancements are use of SSE3 (Streaming SIMD Extensions) technology; 2 MB of L2 cache, compared to 512 KB L2 cache on current offerings; and enhanced efficiency of the Hyper-Threading Technology. Additionally, the new processors will operate using an 800 MHz processor system bus.

Occupying 1,100 square feet, the nominal system is the IBM BladeCenter 2 consisting of 2,048 nodes of two processors each. The CPUs will run at least 4.0 GHz and produce 2 flop/s per clock. This means the peak performance of the processor will be 6.4 Gflop/s and the entire system will be 32.4 Tflop/s. Each node will have 8 GB of memory, and the system will have over 340 TB of usable shared disk. The memory subsystem provides 3.2 GB/s in memory bandwidth per CPU and an interconnect bisection bandwidth of 2 GB/s. The sustained performance of this system is expected to be at least 3.3 Tflop/s.

In order for NERSC leadership users to be most productive, the NCS-L will have a programming environment very similar to that of the NERSC-5/6L. NCS-L and the NERSC-5/6L systems will share a high performance global parallel file system, thus eliminating the need for users to transfer files as they work on different systems. Also, the systems will share a scheduling system for jobs so users can schedule work on any of the systems regardless of which one they are currently working on.

3.4 Petaflop/s Computing in 2010

After numerous discussions with IBM, NERSC believe that it is feasible for NERSC-7P to be a parallel system that evolves from the science-driven architecture process. It will be a large system in the 2010 timeframe that incorporates the best features from the Power series as well as IBM's research projects in Blue Gene, DARPA's HPCS program, and cell-based computing. It will be defined in collaboration with NERSC, IBM, NSF partners, and other DOE research groups. Its cost will be about one-third that of the NERSC-6L system (~$100M for hardware) on a per flop basis and will be one of the

first petaflop/s systems in existence for production-quality science. There is reason to believe price performance may be conservative, since it is based on projects and factors that have been known for some time (BG/L, HPCS/PERCS, Power Architecture, etc.) Other "disruptive" technologies (optical switching, cell-based CPUs, etc.) have the potential to improve performance between a factor of 2 and 5.

IBM is very interested in pursuing this path in collaboration with NERSC.

3.5 Science-Driven Architecture: A Persistent Effort for ViVA-2 and Petaflop/s Development

We propose a program of collaborative design to ensure the continued availability of science-driven computer architectures. Work in FY2005–2006 will focus on ViVA-2 extensions to Power6/6+ processors and systems. ViVA-2 is a science-driven application accelerator, targeting bottlenecks that degrade scientific code performance.

NERSC computational and computer scientists, and NERSC application users will become integral contributors to the IBM development teams for future systems. In collaboration with computer scientists, the applications requirements will be translated into clearly quantifiable measurements of architectural parameters. In ongoing development meetings, the hardware designers will use these requirements to explore design trade-offs and will propose prototype architectures back to the applications teams. Iterating this process will lead to defining prototype designs. IBM will then deliver small-scale early prototypes for actual applications testing. Further iterations and refinements will then lead to the final, large-scale production machine.

In FY2007–2008, work will focus on system enhancements that will result in a petaflop/s system in FY2010 that has been designed for scientific applications. Specific areas of investigation will be determined by experiences with ViVA-2 and the likely configuration of 2010 systems.

The level of effort required to have a major impact on a large commercial system is significant and above what the current NERSC staffing level can support. We propose a level of effort of 10 FTE at Berkeley Lab, collaborating with up to 4 FTE in the IBM design teams, and $500K equipment budget per year. The effort will be divided as follows:

- 5 FTE (NERSC) — Detailed application analysis. This labor-intensive work will focus on evaluating the quantitative impact of architectural changes on the performance of the scientific workload.
- 4 FTE (IBM) — Detailed system design studies for potential features and functions.
- 3 FTE (NERSC) — Simulator and tool development. The tools to perform the detailed application analysis will be quite complex. Work will leverage the SciDAC PERC and DARPA HPCS modeling efforts as well.
- 2 FTE (NERSC) — Senior architects. These individuals will lead the design effort at Berkeley Lab.

Scalable Architectural Prototype (SAP) in 2007

To influence system and processor designs, work has to begin four to five years ahead of system introduction. The work for Blue Planet has been under way at NERSC and LLNL since 2002, and will result in system functionality in the Power5 (2005) and the Power6 (2008).

It is important to prototype and evaluate the leading features for a petaflop/s system for 2010, and doing that in 2007 not only gives sufficient time to influence the 2010 petaflop/s design, but it is also timely due to technology availability. The SAP system will combine experiences from IBM's BlueGene research program and their investigations of cell-based CPUs. IBM has several advanced technology

initiatives that are of mutual interest to NERSC. Specifically, BlueGene/L (power-efficient processors, special-purpose interconnects, high chip-level integration, system-on-a-chip) and Zebra (Playstation-3 game console or "cell" processors with special-purpose streaming units, system-on-a-package). Merging the possibilities of these technologies to achieve performance could provide a dramatic improvement in computational capability. These new initiatives focus on high levels of integration, power-efficient design techniques, and targeted use of special-purpose hardware.

IBM and NERSC plan to create a collaboration in 2005–2006 that would evaluate and integrate these technologies. During 2007 NERSC would target the acquisition of a substantial research version of an advanced technology system from IBM. The system would be built using these components and could yield over 10 Tflop/s peak in a single rack.

There are a number of challenges to achieving good performance for a broad class of applications through these techniques. The programming environment as well as the programming model itself needs to be seriously addressed. There is a realization that techniques to enhance productivity of these emerging technologies in one area may be ineffective in another area. This indicates that the options and costs associated with them need to be deeply understood.

The SAP architectural prototype will explore the best combination of features from Blue Gene/L, cell workstations, and other IBM research investigations. The system will be approximately 100+ peak Tflop/s for a cost of about $20M for hardware. There is the potential that this system, once it is defined, will use such disruptive technology that it will have a performance 2 to 5 times this amount!

NERSC expects this system to work well enough that it will be able to support at least a subset of the application areas running on the production leadership systems. There are definitely some users who can achieve excellent science results while working on prototype systems that are not yet full production. Based on our experience with BG/L and the Power systems, we expect that on these codes, at a minimum, the SAP system will be 10% efficient, yielding about 10 Tflop/s sustained performance. NERSC will push to provide access to as wide a range of applications and users as makes sense for the prototype nature of the system.

NERSC, IBM, NERSC users, and other DOE sites will be able to investigate application performance and scalability issues, operating system scaling, and the ability to deploy a system that is environmentally more friendly.

3.6 System Summary

Table 3 summarizes the systems proposed.

Table 3
System Summary

NERSC Enhancements					
	NERSC-5L	**NCS-L**	**NERSC SAP**	**NERSC-6L**	**NERSC-7P**
Installation Date	2QCY05	1QCY06	2QCY07	1QCY08	~1QCY10
Operational Date	3QCY05	2QCY06	3QCY07	2QCY08	~2QCY10
Processor Type	Power5	BladeCenter 2 with Intel / Irwindale	Cell hybrid	Power6	TBD
Processor Clock Rate (GHz)	1.9	4		5	
Flops / Clock Cycle	4	2		4	
Processor Peak (Gflop/s)	7.6	8		20	
System Peak (Tflop/s)	62.3	32.8	100 or more	327.7	1000.0
System Sustained (Tflop/s)	11.2	3.3	10 or more	78.6	TBD
Efficiency	18%	10%	~10%	24%	
Processors per Node	8	2		8	
Total Nodes	1,024	2,048		2,048	
Total Processors	8,192	4,096		16,384	
Simultaneous Multi-Threading	SMT	HT		SMT	
L1 Cache (KB)	4-way Associative LRU			4-way Associative LRU	
L2 Cache (MB)	Shared across 2 good core processors			Private per core	
Memory	333 MHz DDR1	400 MHz DDR2		1.5 GHz DDRn	
Memory / CPU (GB)	4	4		4	
Memory / System (TB)	32	16		64	
Memory Bandwidth					
Memory BW / CPU (GB/s)	16	3.2		72	
Memory BW / System (TB/s)	128	6.4		576	
Memory BW / Peak Performance (B/F)	2.11	0.40		3.60	
Interconnect					
Technology	Federation	Infiniband 4X or Mryinet 10G	Derived from Blue Gene/L and other R&D projects	Infiniband 12X (DDR)	TBD
Inter-node MPI Bandwidth	4 GB/s X 2 planes	4 GB/s X 2 planes		12 GB/s X 2 planes	
Inter-node MPI (Network) Latency	<5.5	3-5		2.5	
Bisection Bandwidth per flop (GB/s)	0.13	0.13		0.15	
Bisection Bandwidth (TB/s)	4.0	2.1		24.6	
Global Disk					
Technology	Serial ATA	Serial ATA		Serial ATA	
Raw Disk Space (TB)	880	512		4,500	
Usable Global Disk (TB)	640	340		3,300	
Global Disk / Compute Ratio (Bytes/Flop)	10.28	10.38		10.07	
Total Floorspace (Sq. Ft.)	4,800	4,800	~500	9,500	
Electrical Power (System Only) (MW)	3.2	0.62	4.5	7.16	
Electrical Power (System and Cooling) (MW)	4.8	0.927	1.8	10.74	
Bundled Cost ($M)	81.4	28.9	20	123	~100
Maintenance Start	3QCY08	2QCY09	3QCY08	2QCY09	2QCY10
Maintenance Cost ($M)	8.1	2.2	~2	12.3	~10

4. RESOURCES REQUIRED FOR A SCIENCE-DRIVEN SYSTEM ARCHITECTURE

Resources required for a science-driven system architecture are discussed below. These include intellectual resources, infrastructure and capabilities, and building and physical infrastructure.

4.1 Intellectual Resources

Berkeley Lab: A Proven Track Record in Computational Science and Computer Science

NERSC directly benefits from the pool of talent available in the two research departments of the Computational Research Division at Berkeley Lab. NERSC collaborates closely with ESnet to provide high-bandwidth access to NERSC resources. There are also increased opportunities for technology transfer from other DOE/OASCR-funded projects elsewhere at Berkeley Lab, especially the Scientific Discovery through Advanced Computing (SciDAC) programs led by Berkeley Lab PIs.

- **The High Performance Computing Research Department (HPCRD)** addresses long-term research and development questions in HPC. With more than 125 staff and expertise in computer science, computational science, and applied mathematics, HPCRD can provide additional resources and talent for the advanced development needs of NERSC and for focused high-end support of the application areas.
- **The Distributed Systems Department (DSD)** focuses on issues in distributed computing, Grid technologies, networking research, collaborative tools, and security. With more than 25 staff, DSD develops and prototypes technologies and testbeds to facilitate solving scientific problems that require complex and large-scale computing and data handling environments involving geographically and organizationally dispersed components. DSD can provide additional resources and talent for enabling the distributed infrastructure for NERSC applications areas.
- **SciDAC Centers:** Berkeley Lab is the leader of four SciDAC centers and eighteen SciDAC projects. NERSC will leverage the activities of these projects, in particular the Applied Partial Differential Equations Center (APDEC) and Performance Evaluation Research Center (PERC) Integrated Software Infrastructure Centers.
- **ESnet:** To meet the science requirements of the program offices in DOE's Office of Science, ESnet provides high-bandwidth connectivity, guaranteed bandwidth services, and highly reliable network connectivity. ESnet's long-term response to these requirements is a new network architecture that involves connecting the DOE science sites with ring-structured Metropolitan Area Networks (MANs) that are "dual threaded" by more than one national core network. Each national core would connect to the MANs at different physical locations for increased wide-area reliability.

  This architecture provides three major benefits. First, the MANs will be based on multiple 10 Gb/s optical channels ("lambdas") that provide high-speed access to the ESnet core network. Second, the ring structure of the MANs will provide the labs with redundant access to the network, thus providing substantially increased reliability. Third, multiple optical channels will allow for ESnet to provide new services identified in the science requirements, in particular guaranteed high-bandwidth channels.

  The first ESnet MAN will be built in the San Francisco Bay Area and will connect, in a 10 Gb/s ring, DOE's Stanford Linear Accelerator Center, Lawrence Berkeley National Laboratory, the

Joint Genome Institute, NERSC, the ESnet core network hub in Sunnyvale, and the Level3 Sunnyvale site that includes a National Lambda Rail hub that gives access to DOE's Ultra-Science net.

University of California, Berkeley

Berkeley Lab's location, only a short walk or shuttle bus ride away from the campus of the University of California at Berkeley (UC Berkeley), facilitates numerous formal and informal collaborations. Currently, there are seven joint appointments of faculty from the Electrical Engineering and Computer Science (EECS) and Mathematics Departments at UC Berkeley with Berkeley Lab Computing Sciences: David Culler, James Demmel, Susan Graham, Ming Gu, Arie Segev, Jonathan Shewchuck, and Katherine Yelick. Jim Demmel is also the Chief Scientist for the Center for Information Technology Research in the Interest of Society (CITRIS), a four-campus, 200+ faculty research institute centered at Berkeley. The combination of NERSC facilities and Berkeley Lab and campus computing efforts creates a vibrant community for cross-institution and cross-discipline efforts in research in algorithms, architectures, and applications, and in training of future computational scientists.

Katherine Yelick leads the Berkeley Unified Parallel C (UPC) team, a collaborative effort centered at Berkeley Lab, which is working to produce more efficient and productive programming models. Yelick is also working with Berkeley Lab scientists on the evaluation of advanced architectures for scientific computing, including processor-in-memory, streams, VLIW, and vectors. This architecture evaluation team worked closely with IBM in the early stages of ViVA design to understand the benefits and limitations of vectors, and what type of memory system was needed to support the more challenging DOE applications. The close research interactions between the UC Berkeley campus and Berkeley Lab have had tremendous impact on DOE science and technology development thus far.

Lawrence Livermore National Laboratory (LLNL)

LLNL has an existing collaboration with IBM to field the ASCI Purple and Blue Gene/L systems. At their introduction, these will be the most powerful systems in the world. LLNL is looking back over a ten-year history of fielding some of the most powerful and innovative systems, often the first of their type. Berkeley Lab and LLNL have collaborated in many ways in the past, most recently in the design of the 8-way node. LLNL will bring the following elements into the NERSC-L effort:

- Collaboration in standing up and operating the next generation of IBM platforms. With ASCI Purple, of similar design to NERSC-5L, being installed first, NERSC will be able to learn from the LLNL experience. NERSC and LLNL will exchange staff: staff from Berkeley will work side by side with LLNL staff when ASCI Purple comes on line, and vice versa. In the long term, after NERSC-5L is on-line, LLNL and NERSC agree to share operational information, e.g., trouble-tickets, etc.
- Share LLNL's planning documents for storage-area network (SAN) architecture, including I/O Blueprints. We will continue to work together with the High Performance Storage System (HPSS) consortium, using our collective leverage with IBM and our combined HPSS development staffs to assure that the appropriate solutions are rapidly written into the HPSS releases.
- Share, test, debug and deploy together the latest ASCI tools in visualization, including utilization of commercial technologies to achieve new levels of graphics performance, the distributed parallel rendering software stack (Chromium), parallel, scalable end-user applications (like VISIT and Blockbuster movie player) and the blueprint for future Purple 100 TF-related visualization deployment.
- As members of the BlueGene/L (BGL) Consortium, NERSC will work together with LLNL to evaluate the appropriateness of the BlueGene/L and the BG-family (BG/P follow-on architecture)

as a leadership-class investment later in this decade by the Office of Science. BlueGene/L represents a $100M R&D investment by IBM in a machine for science, and employs three separate networks to enhance efficiency and an extremely low power system on a chip design. The results of this shared evaluation effort will likely drive changes in both the BG/P and NERSC-6L designs and will have significant importance in defining the road to petaflop/s. IBM will make available to NERSC the LLNL SLURM and LCRM fair-share scheduling and node-packing software, should NERSC choose to employ this solution rather than native software.

- Staff from LLNL who are active in the ASCI program will be part of the quarterly progress meetings that NERSC will have with IBM.

4.2 Infrastructure and Capabilities

Networking

Berkeley Lab and NERSC are located near the primary switching point for national networks in Northern California at Sunnyvale — home to both the Qwest and Level3 networking hubs. The Qwest hub is the transit point for the backbones of major production networks such as DOE's ESnet, NSF's Abilene, NASA's NREN, and the NSF TeraGrid, while the Level3 hub carries experimental dark-fiber networks such as the National Lambda Rail, the DOE Ultranet, and the CENIC/Pacific Light Rail. The proximity allows NERSC easy and cost-effective access to each of these networks. In order to promote interaction with and outreach to scientists in industry, academia, and other federal programs, NERSC will work closely with ESnet to create network peering arrangements that will maximize the effective remote access to NERSC users regardless of their institutional affiliation and facility location.

In order to ensure the highest performance network access, NERSC will immediately upgrade its connection with Sunnyvale to OC-192 in order to match the existing backbone bandwidth of the ESnet and Abilene production networks. In order to provide more effective access to the NSF user community, ESnet and Abilene are implementing high-speed peering between their networks at each of these co-located hubs at Sunnyvale, Chicago, New York, and Atlanta to create a common network backplane that provides very high-speed connectivity between the labs and universities, comparable to what either backbone alone can provide among their primary sites.

In close collaboration with ESnet and CENIC, NERSC will join the Bay Area Metropolitan Area Network (MAN) about the time NERSC-5L arrives. This 10 GB/s link will greatly enhance access to NERSC from all the major networks. Because this MAN is constructed from dark fiber, it will be feasible to add more bandwidth as NERSC needs it for little or no additional cost.

In addition to its support of production network infrastructure, the NERSC system will consider connections to major experimental and dark fiber networks, such as the TeraGrid, DOE Ultranet, and National Lambda Rail, in order to add its capabilities to a vibrant research community that combines sensors, archival data, and supercomputers to accomplish large multidisciplinary scientific projects. Both the upgraded internal network and wide area network infrastructure will be immediately available to the NERSC-5L system and will continue to be expanded to match the scale of successive systems and continuously match the performance improvements of the production network backbones.

In the first year of NERSC-5L operation, ESnet will deploy an MPLS-based QoS service that operates initially between ESnet border routers which will be expanded within two years to allow dynamic provisioning of circuits across both Abilene and ESnet as envisioned by the Internet2 Hybrid Optical/Packet Infrastructure (HOPI) working group. These "bandwidth corridors" will support NERSC global file system (WAN GPFS) and storage peering arrangements between other laboratories and collaborator sites such as the National Science Foundation's Partnership for Advanced Computational Infrastructure (NSF-PACI) supercomputing centers in order to support our vision of a nationwide supercomputing infrastructure.

Systems Management

NERSC will draw on its expertise with past systems to customize the support model for the leadership-scale systems to be more tightly integrated with the selected science projects. NERSC already manages three distinct systems with different user communities and requirements.

The NERSC leadership-scale systems will be operated in a dynamic manner in close collaboration with the users. NERSC will provide a custom and flexible environment that supports the unique requirements of each project — not just for system management but for all support functions. For example, libraries and middleware will be selected and installed in close collaboration with the projects. Users will be able to request to use large amounts of the resource interactively for debugging and computational steering. At times, it will be possible for a user to be given the entire system in a dedicated manner. Because there is a smaller, more manageable set of projects, NERSC staff will be able to coordinate the system scheduling to meet computational science project goals in a custom manner. NERSC will involve users in the discussion of system management changes — in particular queuing, priorities, and disk-space management — through monthly conference calls.

Data Storage and Archives

NERSC's High Performance Storage System (HPSS) will continue to have enough capacity to serve all of NERSC's clients. NERSC currently stores approximately 1,050 TB of data (30 million files) and handles between 3 and 6 TB of I/O per day. The current maximum capacity of NERSC's archive is 8.8 PB at current tape densities; the buffer (disk) cache is 35 TB; and the maximum transfer rate is 2.8 GB/s. The leadership-class systems will require large amounts of archival storage, and NERSC will invest in new tape technology. For NERSC-5L, 500 GB tape drives and cartridges will be added to the NERSC HPSS, giving a total maximum capacity of 4.5 PB just for NERSC-5L. For NERSC-6L, 1 TB cartridges will be deployed, adding 5 PB a year (10 PB total for the time period of the proposal) to the potential NERSC storage capability, for a total of 30 PB of storage.

The NERSC HPSS will be federated with archival storage systems (both HPSS and Unitree) at sites used by NERSC-L users, who will have equal access to archival data across NERSC, NSF-PACI facilities, and ORNL through a storage federation. Close coordination of certificate management between DOE Science Grid, TeraGrid, ORNL, and NSF-PACI sites will enable single-sign-on access across facilities and seamless transfer of data between archival storage systems. Also, the bandwidth corridors described above will support dedicated high-speed data transfers between the sites for efficient mirroring and staging of massive datasets between their respective storage systems. NERSC will work in particular with DOE sites such as ORNL to make it as easy as possible for scientists who use both archive systems. The effort will include integrating archive data transfer with site security policy, optimizing transfer tools to move data using multiple nodes on systems, and implementing multi-stream transfers for improved performance. The concept of this work was proven several years ago in the prototype PROBE storage testbed, run jointly by NERSC and ORNL.

In addition to archival storage systems, NERSC will be part of a wide-area shared file system that will link together partner sites including the NSF-PACI supercomputing centers and other DOE laboratories. The file system will be based initially on WAN GPFS, which is being developed through a partnership between IBM Research and the San Diego Supercomputer Center (SDSC), and will be usable across both Linux and IBM SP supercomputing infrastructure at participating sites. In demonstrations conducted by SDSC this past year, GPFS sustained well over 900 MB/s over a wide-area 10 gigabit link. The shared file system will enable more flexible migration between the systems for users who have shared accounts and will help the leadership collaborations form a well-integrated computing environment that better serves a national scientific user community

Grids

As the home of ESnet and NERSC, the lead site for the DOE Science Grid, and one of the original six development sites for the HPSS, Berkeley Lab has already made significant progress in integrating high-end computing, storage, and data management into the Grid environment. We will continue facilitating large-scale science for DOE and the nation by extending this technology and expertise. NERSC has established ties with all major Grid efforts in DOE and many in the NSF and is closely collaborating with the DOE Science Grid and all its partners. Because NERSC supports a wide range of Grids and Virtual Organizations, it plays a unique position in the Grid effort — being a unique Grid "hot spot" where many individual grids overlap. The NERSC center staff leverage its broad experience with Grid software and services. We will work in close coordination with the other sites to establish the peering of Certificate Authorities and trust relationships necessary to support coordinated access to Grid services. An interface to the NERSC Information Management (NIM) system makes it easy for NERSC users to get Grid authentication certificates. Coordinated management of Grid certificates supports single-sign-on access to Grid services across partner sites including the NSF-PACI centers, NERSC, PNNL, other DOE laboratories, and the TeraGrid Consortium.

Visualization and Data Analysis

High-end visualization and data analysis tools will be essential to turn raw simulation data into scientific discoveries. NERSC works closely with its partners to apply technologies developed across the coalition and make them available to the user community. In particular, we will work closely with LLNL to share, test, debug, and deploy the latest ASC tools for visualization of massive datasets, including commercial technologies that offer new levels of graphics performance, the LLNL/Stanford-developed distributed parallel rendering software (Chromium), and proven parallel, scalable end-user applications (like VISIT and Blockbuster movie player), and the Terascale Browser. The Berkeley Lab/NERSC visualization group will also provide NERSC users with access to the VisPortal, which automates complex workflows like the distributed generation of MPEG movies or scheduling of file transfers, mediates access to limited hardware resources like off-screen graphics pipes, and controls the launching of complex multicomponent distributed visualization applications like Berkeley Lab's Visapult — an application used for remote and distributed, high performance interactive volume rendering of massive remotely located datasets. All of these tools will be tightly coupled with the high-speed networks, coordinated Grid services, storage federation, and WAN GPFS capabilities deployed across the sites. This powerful set of tools and services will enable users across the nation to rapidly understand the enormous amount of data they generate at NERSC. Without tools of this caliber and computer scientists available to support these tools, the huge data generation engines that NERSC will be deploying would be less useful.

Security

As an unclassified facility, NERSC makes its facilities available for use by investigators from institutions throughout the nation and the world. To sustain its scientific mission, NERSC protects its resources and assets, both intellectual and material. Only necessary technical staff have access to computer rooms and computer facilities. The general staff and the public do not have physical access to these computer resources. All facility assets are tracked and protected by Berkeley Laboratory security services. NERSC users will access the system remotely, subject to all Berkeley Lab cyber security policies, controls, and restrictions.

Berkeley Lab and NERSC have an outstanding security record and are recognized as leaders in cyber security within the DOE and beyond. In fact, during the recent serious cyber attacks that disrupted service at many supercomputer facilities throughout the US, NERSC suffered no system compromises and no service disruption.. This expertise makes NERSC both secure and easily accessible. In order to maximize our ability to conduct science and mitigate the effects of computer security incidents, NERSC provides non-invasive advanced monitoring and automatic reactive tools using components that are embedded in

the network as well as in every computational and storage system. NERSC's active security infrastructure is able to detect cyber attacks, detect vulnerable or compromised hosts, and initiate a large-scale coordinated response to cyber-security incidents without resorting to methods that impede legitimate system access. For example, firewalls are creating significant roadblocks to pervasive deployment of Grids and high-bandwidth networking. Berkeley Lab's active intrusion detection system, Bro, offers a compelling alternative to standard firewalls as a means to defend against cyber attacks. DOE is funding efforts to extend this system to sites other than Berkeley Lab. The Laboratory will continue to use and improve these advanced monitoring tools to provide NERSC with the best level of security with minimal impact on performance and function.

4.3 Building and Physical Infrastructure

Berkeley Lab's Oakland Scientific Facility (OSF) includes a 20,000-square-foot computer floor. Currently, there are 5,500 square feet of computer floor available for additional systems. The NERSC-5L, described in this proposal, will require 4,800 net square feet and will readily fit in the existing OSF. Then NCS-L and Architectural Prototype systems will fill in the remaining space in OSF.

The follow-on system, NERSC-6L, will be housed in a new computer building in the center of the Berkeley Lab campus on a cleared site adjacent to the Bevatron, whose external beam hall was recently demolished. The building will contain a computer room and utility support space. This cleared site also provides for the ability to expand into a second adjacent 20,000 square feet of computer floor, yielding a 40,000-square-foot computer complex, thus allowing all of NERSC to locate to the new central facility. Site plans and conceptual building renditions are shown in Appendix E of [2].

Recently, the DOE has encouraged third-party financing approaches to facilities construction, and these approaches will enable Berkeley Lab to provide the requisite leadership computing building to accept delivery of NERSC-6L. Because Berkeley Lab is located on University of California-owned land, this process is actually less complicated for Berkeley Lab than for those national laboratories situated on federal land, which must be transferred via a quitclaim deed to a development entity. The University can simply enter into a long-term ground lease with a developer at a nominal cost. When the building is complete, DOE can approve a UC lease of the facility a year at a time over the life of the building. Berkeley Lab and the University are currently developing a research office building on the main Lab campus, targeted for completion in 2006, through a third-party development. The experience and knowledge gained from this procurement give us every confidence that the computing building can be completed on time.

The contingency plan for housing NERSC-5L and NERSC-6L is expansion of the OSF to gain another 20,000 square feet. The OSF was designed for such a contingency, which can be exercised in time for NERSC-6L.

Berkeley Lab, therefore, has existing and committed space for leadership-class computing systems.

5. COMMUNICATIONS AND OUTREACH

NERSC, as potentially the largest open computing resource in the nation, has developed collaborations with computational scientists in universities, research labs, and industry. In order to maximize the dissemination of information, and to promote and support computational science and computer technology for high-end computing, NERSC has far-reaching plans for collaborations, outreach, and dialogue with stakeholders. In the areas of technology development, NERSC will engage its major vendor partner, IBM, in an ongoing dialogue of science-driven architecture development. NERSC will work to build on close connections and strategic collaborations with computer science programs and facilities funded by DOE-SC, DOE-NNSA, NSF, and NASA, as well as universities.

The following events will facilitate this outreach:

1. *Monthly meeting/conference call with users*. Leadership systems users will hold a monthly conference call with NERSC staff to discuss operational issues, progress towards system and software deliverables, applications porting and performance issues, etc.
2. *Semi-annual progress meetings*. NERSC staff, representatives from other sites, vendor partners, and application scientists will meet semi-annually to report on progress with their tasks. The semi-annual progress meeting will also serve as the main communication mechanism for the implementation of the science-driven architecture development.
3. *Annual "all-hands" meeting.* NERSC will organize an annual event that will be open to all stakeholders and the community at large. It will include scientific presentations from NERSC users, updates from the vendor partners, and computer science and technology presentations from the NERSC staff.
4. *Workshops and planning meetings.* As new and important topics arise, NERSC will hold workshops and planning meetings for interested stakeholders.

5.1 NERSC Education and Workforce Development

NERSC will develop a leadership computing community through integrated educational and training components that build skilled computational scientists, with a focus on graduate and undergraduate students. Outreach to underrepresented students will be integral to building the educational pipeline. The education program will include:

- Seminars on leadership computing capabilities targeted to specific research topics
- Short courses on specific computational topics
- Consulting services, including course assistance to ensure up-to-date user information
- Web site resources with comprehensive technical, information, and course content
- Internships available to qualified applicants for summer and semester appointments
- Graduate and undergraduate research on all phases of leadership computing
- Faculty sabbaticals to update computing courses and curriculum

A key benefit of NERSC is the combined educational resources of collaborators, such as the internationally recognized education program in computing science and applied and computational mathematics at UC Berkeley. The NERSC ties to the University of California and the NSF-PACI supercomputing centers (SDSC and NCSA) bring national university educational resources to bear on training and future workforce development. NERSC will provide training and internships for the DOE Workforce Development for Teachers and Scientists program.

6. MANAGEMENT PLAN

NERSC is managed as a national scientific resource with full and complementary support to the programs of the Office of Science and mission of the U.S. Department of Energy. Facility management will focus on sound annual planning, cost-effective line management, comprehensive review, and a highly consultative management advisory framework. The management system will be coupled to the Office of Science program evaluation process for program oversight and Laboratory management. The management efforts will reinforce NERSC's mission of demonstrating continued U.S. leadership in computational science through performance at the largest scale of computational problems. NERSC-L system deployment will be planned and implemented though a project management framework to assure the completion of facilities components on schedule, scope, and budget in a manner that is consistent with all affected stakeholders.

6.1 NERSC Management and Organization

NERSC is led by Horst Simon, Associate Laboratory Director for Computing Sciences and NERSC Center Division Director. As NERSC Director, he is accountable for all aspects of the NERSC program. The Director recommends strategic programmatic directions and the development of programmatic ties to other laboratories, universities, and industrial partners, as appropriate. The Director is responsible for scientific productivity and maintaining the leadership role of the facility.

Division directors at Berkeley Lab are direct appointees of the Laboratory Director and are members of the Director's planning team, participating in the Laboratory's oversight and review activities. NERSC has direct access to the Laboratory Director and high visibility at the Laboratory. The organizational arrangement for NERSC is similar to that of the Advanced Light Source, the National Center for Electron Microscopy, the Molecular Foundry, and ESnet, the other major national user facilities at Berkeley Lab.

The NERSC Director is supported by a management team led by a General Manager, who works with other NERSC staff to carry out their responsibilities. The General Manager, William Kramer, reporting to the NERSC Director, is accountable for the NERSC Center, with management responsibility for planning, budgets, enhancements, personnel, vendor and user relations, physical resources, and program and operational integration.

6.2 National Oversight and Policy

NERSC management will meet with leadership of the Office of Advanced Scientific Computing Research (OASCR) and with the Mathematics, Information and Computing Sciences (MICS) program management staff to develop program plans and budgets and to facilitate periodic OASCR and other national reviews of the NERSC program.

The Berkeley Laboratory Director conducts an annual review of NERSC, which includes an assessment of NERSC's long-range planning, staffing, quality of programs and operations execution. The review is conducted by the NERSC Policy Board, which is appointed by the Laboratory Director in consultation with OASCR. The Board consists of leading representatives of the high performance computing community, and provides advice on strategic issues and policy directions to both the Laboratory Director and the NERSC Director. The current members of the NERSC Policy Board are shown in Appendix B.

6.3 Allocation Review Process

In 2003, DOE initiated a new program entitled Innovative and Novel Computational Impact on Theory and Experiment (INCITE) at NERSC. INCITE awarded 4.9 million supercomputer processor hours and corresponding data storage space at NERSC to three computationally intensive large-scale research projects, with no requirement of current DOE sponsorship. A peer review process for all

proposals was established, which involved a Web-based proposal submission system, a review panel of about 95 scientists, and a well-defined process for evaluating both the scientific goals and the computational methods and techniques of the proposal. The NERSC leadership system allocations process will evolve from a combination of the successful INCITE program and the current NERSC allocation process, leveraging both infrastructure and the reviewer pool. The existing Energy Research Computational Application Program (ERCAP) and NERSC Information Management (NIM) systems will be used to implement the mechanics of the allocation process.

7. BUDGET

The budget proposed for this project will achieve 100+ Tflop/s aggregate sustained performance in 2008 and define and field the petaflop/s system in 2010. The costs presented here represent the total cost of ownership for the six-year life of the project. They include all staffing costs (salary, benefits, burden and support), all system costs (hardware, software, maintenance, space, procurement burden, electricity for running and cooling the systems, and space for housing the systems), and the infrastructure to connect NERSC-L systems with the NERSC and Bay Area MAN infrastructure.

The infrastructure to connect the NERSC-L systems to NERSC's infrastructure consists of two major parts:

1. $7.8M to provide archive storage capacity and bandwidth proportional to leadership-class capabilities. This is approximately 30 PB of storage by the end of 2009. This leverages NERSC HPSS services, software, and caching disk, and provides only the additional tape drives and tapes to hold data.
2. $5.0M to provide improved high bandwidth networking to be commensurate with the power of the systems. This includes the switch and enough network interfaces for all computing and storage systems, but again leverages NERSC significantly.

Table 4 shows the overall costs of the proposal, which total $513M. NERSC proposes to contribute $44M over six years from its base program of $28.244M per year, so the final enhancement cost is $469M. The breakdown of the funding is:

- 92.2% of the funding is for ownership of the leadership computational systems (NERSC-5L, NCS-L, SAP, NERSC-6L, and NERSC-7P).
- 2.9% of the cost is for the archive and network storage directly needed by the leadership systems.
- 4.9% is for staff providing a persistent science-driven computer architecture effort that will assure the success of the advanced features discussed here.
- No direct funding is needed for the new facility, since it is being provided by the Laboratory.

Figure 2 shows a spending plan without constrained funding. It shows a non-uniform investment pattern that shifts from year to year. NERSC will work with the DOE to justify a non-uniform funding profile, use mechanisms to carry over funding, or use lease-to-own arrangements to even out the cash flow of the solutions. The breakdown of capital-to-operating funding will be set based on the final configurations and arrangements.

The budget is flexible, and NERSC is open to feedback and input. For example, it may be desirable to have a larger NERSC-5L system, which could be achieved by either adding more funding or decreasing the size of the NCS-L cluster.

Table 4
Budget Summary

Budget Summary

Budget Summary (dollars in thousands)		FY 05	FY 06	FY 07	FY 08	FY09	FY10	
SDCA Personnel (in Full Time Equivalents, FTE)								
Application Analysis	5							
Simulator and Tool Development	3							
Senior Architects	2							
System Design	4							
Total FTEs	14							
Staff Costs								
Direct Salaries		$1,578.2	$1,625.5	$1,674.3	$1,724.5	$1,776.3	$1,829.5	
Burdens		2,220.5	2,287.1	2,355.7	2,426.4	2,499.2	2,574.2	***TOTALS***
Other Support (Travel, etc)		88.3	90.1	91.9	93.8	95.7	97.6	***in $(000s)***
Additional Staff Costs		$3,887.0	$4,002.8	$4,122.0	$4,244.7	$4,371.1	$4,501.3	$25,128.9
Systems Costs								
Computational Investment, Maintenance & Facilities (incl. electricity)		$83,012.0	$35,276.1	$25,885.4	$145,175.3	$38,055.6	$145,791.5	$473,195.9
Network		$659.6	$1,209.8	$415.4	$473.2	$620.7	$850.8	$4,229.5
Storage		$1,997.3	$888.8	$1,178.1	$669.8	$1,669.8	$669.8	$7,073.7
SDSC Test Bed		$558.2	$558.2	$558.2	$558.2	$558.2	$558.2	$3,349.2
Total Systems Costs		$86,227.1	$37,932.9	$28,037.1	$146,876.5	$40,904.3	$147,870.3	$487,848.2
Enhancement Costs		$90,114.1	$41,935.7	$32,159.1	$151,121.2	$45,275.4	$152,371.6	$512,977.1
NERSC Base Program Contributions*								
Contributions from the NERSC Base Program		$9,800.0	$4,359.0	$3,000.0	$10,000.0	$10,239.0	$6,708.0	$44,106.0
Grand Total		$80,314.1	$37,576.7	$29,159.1	$141,121.2	$35,036.4	$145,663.6	$468,871.1

* Base program assumption is $38M in FY 05 and $28.24M for other years.

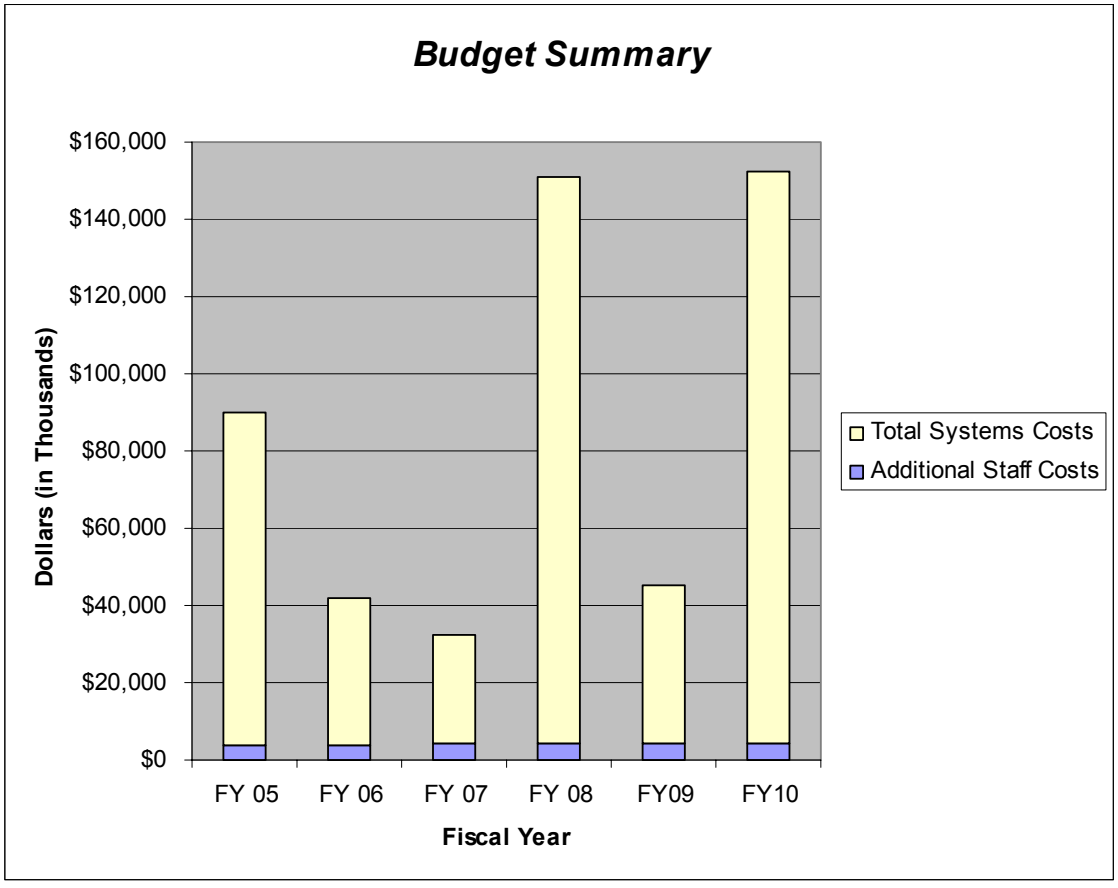

Figure 2. More than 95% of the funding is going to the hardware/software systems.

8. SCHEDULE

NERSC-L Semi-Annual Progress Meetings

07/2005 Kickoff meeting with users
10/2004 Semi-annual meeting
03/2005 Semi-annual meeting
... will be scheduled every six months through
10/2010 Semi-annual meeting

Science-Driven Computer Architecture Milestone

12/2005 ViVA-2 Design complete

Systems

06/2005 NERSC-5L Installation
09/2005 NERSC-5L First User Access
11/2005 NERSC-5L Full user access
03/2006 NCS-L Installation
06/2006 NCS-L First User Access
09/2006 NCS-L Full User Access
03/2007 NERSC SAP Installation
03/2008 NERSC-6L Installation.
06/2008 NERSC-6L First user access
11/2008 NERSC-6L Full User Access
03/2010 NERSC-7P Installation
09/2010 NERSC-7P First User Access

Contracts

03/2005 Detailed statement of work with IBM for NERSC-5L
10/2005 Detailed statement of work with IBM for NCS-L
02/2007 Detailed statement of work with IBM for NERSC-6L

APPENDIX A
Performance Analysis of NERSC-5L and NERSC-6L Systems

This appendix presents a quantitative analysis of NERSC-5L/NERSC-6L performance.

Processor Overview

Key parameters of the Power processors are given in Table A-1. Power5 and Power6 demonstrate significant improvements in memory bandwidth relative to their predecessors, on par with vector machines.

Table A-1
Specifications for Power Processors

	Power4	Power5	Power6
Processor clock rate	1.3 GHz	1.9 GHz	5 GHz
Floating point operations per cycle	4	4	4
Processor peak Gflop/s	5.2 Gflop/s	7.6 Gflop/s	20 Gflop/s
Processors per node	32	8	8
Memory bandwidth per processor	5.2 GB/s	16 GB/s	72 GB/s
Memory bandwidth to peak performance (bytes/flop)	1	2.1	3.6
Average sustained performance (% of peak)	9.5% (single) 6.7% (multi)	15%	24%

Average sustained performance is discussed in detail in subsequent sections. This table demonstrates that IBM has made a substantial investment in processor technology to meet the needs of the HPC community, approximately doubling the effectiveness of their processors. Much of the Power5 improvement is due to the Blue Planet collaboration, which we are continuing and expanding in this proposal. Power6 performance is a baseline, not taking into account the planned further enhancements that are discussed in this proposal.

Projection Methodology

An increase in clock speed does not necessarily yield an increase in sustained processor performance, unless memory bandwidth increases proportionally. When projecting performance increases between processors in the same family, we take into account clock speed and memory bandwidth. If relative memory bandwidth, measured in bytes/flop, remains the same, sustained performance as a fraction of peak should remain the same. If relative memory bandwidth increases or decreases, sustained performance as a fraction of peak will increase or decrease.

Power5 Test System Description

We have had access to an early implementation of a Power5 system. The system is similar to the NERSC-5L system except for the following points:

- Processor speed of 1.6 GHz (NERSC-5L will be 1.9 GHz).
- Two processor cores on a chip. The NERSC-5L system will have one core per chip. To simulate the NERSC-5L system, we did not use the second processor core.

- Compiled using Power4 technology — no Power5-specific optimization.

Projection to NERSC-5L Power5 System

- The processor clock of the NERSC-5L system is 19% faster than that of the test system, and relative memory bandwidth increases by 25%. When increasing memory bandwidth and processor clock speed by the same factor, sustained performance as a fraction of peak will remain the same. When increasing memory bandwidth relative to clock speed, the rule of thumb is that the actual performance increase is about 50% of the relative increase in memory bandwidth. This methodology yields a performance projection of 18.2% sustained performance relative to peak for NERSC-5L.
- The Power5 has some features, particularly a new instruction to start the prefetch engine, that were not targeted by the Power4 compiler we used. We expect a 3% improvement when the new compiler comes out. We believe that this is a conservative estimate, given that that targeting Power4 vs. Power3 yields an 18% improvement in performance.

Combining these two factors yields sustained performance of 21% of peak. Being conservative, we project an average sustained performance of 15% of peak on the NERSC-5L Power5-based system.

Projection to NERSC-6L Power6 System

Going from Power5 to Power6, the improvements include:

- Increase in memory bandwidth from 2.1 bytes/flop to 3.6 bytes/flop (factor of 1.7). Using the methodology above projects sustained performance of 35% for NERSC-6L.
- Introduction of quad-word load into the instruction set (increases the effective number of memory operations “in flight”).

Being conservative, we project an average sustained performance of 24% of peak for NERSC-6L.

Average Sustained Performance

To get a baseline assessment of Power5 performance, we ran the NAS Serial Benchmarks, v3.0, class B. The six codes in this suite represent a variety of numerical algorithms from the field of computational fluid dynamics and are described in [A1]. The results for the Power4 and Power5 test systems are shown in Table A-2.

The ratio of absolute performance of the Power5 test system to the Power4 system is 2.2 (1035/477) while clock speed increased only by a factor of 1.23 (1.6/1.3). This table also shows that a single processor of the NERSC-5L system should perform 3 times faster on average than a Power4 processor. When corrected for memory contention on multiprocessor runs (see below), this performance advantage rises to 4.36.

Application Benchmarks

We were able to run several representative single-processor application codes on the Power5 test system. Results are presented in Table A-3. Run times for each code are given in seconds. We note that the average performance for this broad range of applications increases by 2.3x compared to the Power4.

Table A-2
Performance of NAS Benchmarks

	Power4	Power5 (Test System)	NERSC-5L (projected)
Clock rate (MHz)	1300	1600	1900
Peak performance (Mflop/s)	5200	6400	7600
NAS Codes (Mflop/s)			
BT	827	1400	2056
CG	113	208	306
FT	514	1060	1557
LU	554	1387	2037
MG	430	1321	1940
SP	426	834	1225
Average (Mflop/s)	477	1035	**1520**
% peak	9%	16%	**20%**

Table A-3
Performance Comparison for Selected Applications

	Power4	Power5 (test system)	Power5 (test system) to Power4 speedup
Cactus	3783	1472	2.6
Chombo	396	252	1.6
Paratec	8936	3843	2.3
SuperLU	193	104	1.9
TLBE	4243	1092	3.9
WRF	3600	2003	1.8
Average:			**2.3**

Descriptions of the codes follow.

- **Cactus**: An astrophysics application that evolves Einstein's equations following the Theory of General Relatively. The 4D formulation (three spatial and one temporal dimension) solves coupled nonlinear hyperbolic and elliptic equations containing thousands of terms; thus making it run efficiently on both scalar and vector systems.
- **Chombo**: Chombo is a framework for implementing finite-difference methods that solve partial differential equations using block-structured adaptive mesh refinement (AMR) methods refined rectangular grids. This benchmark examines performance of a Poisson elliptic solver using the Chombo framework. The calculations on individual grids of an AMR simulation will benefit from vectorization; however, the nonvectorizable calculations, such as pointer-indirection, clustering algorithms, and dynamic load redistribution, will dominate the computational costs for large-scale calculations (Amdahl's law).
- **Paratec**: A materials science applications that performs first-principles quantum mechanical total energy calculations based on Density Functional Theory. The code spends most of its time in vendor-supplied dense linear algebra (BLAS3) and 3D fast Fourier transform (FFT) calculations,

and therefore will generally obtain a high percentage of peak processor performance across different platforms. A network with full bisection-bandwidth is necessary for achieving high performance on large systems, due to the global communication requirements.

- **SuperLU**: SuperLU is a general purpose library for the direct solution of large, sparse, nonsymmetric systems of linear equations on high performance machines. Sparse numerical codes such as SuperLU are a critical component of future high performance computing; however, these methods are at odds with vector architectures, as they characterized by control irregularity, resulting in potential loop-carried dependencies that inhibit efficient data-parallelism.
- **TLBE**: Thermal Lattice Boltzmann Equation. This fusion code performs a 2D simulation of high-temperature plasma using a hexagonal lattice and the BGK collision operator. TLBE is a computationally intensive code, which performs sweeps through a regular 2D grid with static communication along the boundary values; making it well-suited for both scalar and vector architectures
- **WRF**: Weather Researching and Forecasting Model. State-of-the-art weather forecasting code. We expect this code to be well suited for vector architectures.

These results are consistent with the NAS Parallel Benchmarks (NPB) experiments, demonstrating that the improved memory bandwidth of the Power5 test system results in higher sustained performance compared with the Power4 system. Additional increases in application performance are expected when the Power5 microprocessor and associated software reaches maturity in late 2004. Furthermore, it is important to note that only a subset of these applications are expected to perform well on vector architectures. In particular, two emerging computational methods, AMR and sparse matrix computations, are better suited for superscalar-based architectures. Our proposed system therefore offers the most flexible solution, allowing the efficient computation of both established and evolving numerical approaches.

Memory Contention Considerations with Multiple Processes per Node

An important concern with the use of symmetric multi-processor (SMP) systems as building blocks of large computers is memory contention within an SMP node. The per-processor performance of parallel applications is typically less than that of corresponding serial applications because of parallel inefficiencies (e.g., Amdahl's law), but also because of memory contention within a node. This has been a particular concern on Power4 systems, which are based on a dual-core design in which two processors share the same interface to main memory, effectively halving the bandwidth. Power4 systems therefore perform particularly poorly on parallel applications — more poorly than one would expect based on single-processor benchmarks.

An estimate of the effect of memory contention can be obtained by running multiple simultaneous copies of a serial benchmark, and comparing their performance to that of a single copy on an unloaded machine. If there is no contention, performance is the same. We define a benchmark *NPB, which consists of running N-simultaneous copies of each NPB benchmark application on an N-processor system. This can be seen for the Power4 in Table A-4:

This result is consistent with the earlier statement that increasing peak performance without increasing memory bandwidth typically improves performance by half the increase in peak. An analysis based on this rule of thumb predicts 6.9% efficiency.

The NERSC-5L and NERSC-6L systems minimize the effect of memory contention through the following mechanisms:

- Dedicated memory system for each processor, including on-chip memory controller.

Table A-4
Effect of Memory Contention on the Power 4

	Power 4 (single copy)	Power 4 (8 copies in 8-processor partition)
NAS Codes (Mflop/s)		
BT	827	682
CG	113	56
FT	514	345
LU	554	357
MG	430	333
SP	426	319
Average	477	**349**
% peak	9.2%	**6.7%**

- "Single core" design. Other IBM systems have two processor cores on a chip. These processors share cache bandwidth and main memory bandwidth, effectively halving the memory bandwidth per processor.
- Small node design. By having fewer processors in an SMP, the memory interconnect is greatly simplified.
- Processor affinity. The scheduling system ensures that process memory is local to the processor on which the process is running.

We expect the effect of memory contention to be minimal in both the Power5 and Power6 systems. We note that the Power5/6 design is similar to that used in the Cray X1, which also has minimal memory contention. Taking into account memory contention, we therefore expect that NERSC-5L processors (1.44 Gflop/s average performance) will on average exceed the performance of Power4 processors (349 Mflop/s average performance) by a factor of 1520/349 = 4.36.

Networks

The network in the NERSC-5L system is a "dual-plane" configuration of the IBM Federation switch. Each of two links is capable of 4 GB/s bidirectional communication (i.e., 2 GB/s simultaneously in each direction). The total of 4 GB/s of bidirectional bandwidth is shared among 8 processors in a node, for a rate of 1 GB/s.

Federation topology is a similar to fat-tree topology, and provides full bisection bandwidth. A network connecting N components is said to have full bisection bandwidth if any N/2 components can communicate simultaneously with the other N/2 components without interference. The total rate of such communication is bisection bandwidth, and the rate seen by each individual component is bisection bandwidth per processor. Bisection bandwidth per processor in the NERSC-5L system is 1 GB/s.

Latency in the Federation switch will be significantly better than the latency of previous IBM switches. With software improvements coming in Q404, latency measured at the MPI level will be 5 microseconds.

The network in the NERSC-6L system will be based on the industry standard InfiniBand interconnect. Depending on availability of quad data rate InfiniBand, it will either be 2 rails of quad data rate 12x InfiniBand or 4 rails of dual data rate 12x InfiniBand. In either case, bandwidth will be 24 GB/s

per node, or 3 GB/s per processor. The NERSC-6L network will also have full bisection bandwidth, or 3 GB/s per processor of bisection bandwidth.

Message latency in the NERSC-6L network is expected to be 2.5 microseconds.

Conclusion

We expect the NERSC leadership systems and Cray's leadership systems to perform similarly overall, albeit on different ranges of the computational science spectrum. Improved memory bandwidth in the systems will increase average sustained performance. The complementary advantages of the NERSC-6L solution are:

- excellent price-performance
- consistent performance across a range of applications
- ViVA enhancements for Power6 constitute the first step toward a new science-driven architecture
- potential to take advantage of IBM's broad technology research portfolio.

In addition, NERSC-6L complements hardware investments made elsewhere in DOE by providing an alternative architecture route, thus mitigating risk and increasing architectural diversity.

APPENDIX B
NERSC Policy Board

Daniel Reed (Chair)
The University of North Carolina at Chapel Hill
147 Sitterson Hall, Campus Box 3175
Chapel Hill, NC 27599

Albert Narath (Retired)
1534 Eagle Ridge Drive, NE
Albuquerque, NM 87122

Robert J. Goldston
Director, Princeton Plasma Physics Laboratory
P.O. Box 451, Mail Stop 37
Princeton, NJ 08543-0451

Robert D. Ryne
Lawrence Berkeley National Laboratory
One Cyclotron Road, MS-71J
Berkeley, CA 94720

Stephen Jardin
Princeton Plasma Physics Laboratory
P.O. Box 451, Mail Stop 27
Princeton, NJ 08543-0451

Tetsuya Sato
Earth Simulator Center Director-General
Japan Marine Science & Technology Center
3173-25, Showa-machi, Kanazawa-ku
Yokohama-City, Japan 236001

Sid Karin
Professor of Computer Science and Engineering
University of California, San Diego
9500 Gilman Drive, MC 0505
La Jolla, CA 92093-0505

Stephen Squires
Vice President and Chief Science Officer
Hewlett-Packard Laboratories
1501 Page Mill Road, MS 3U-10
Palo Alto, CA 94304-1126

William J. Madia
Executive Vice President of Laboratory Operations
Battelle
505 King Avenue
Columbus, OH 43201

Michael Witherell, Director
Fermi National Accelerator Laboratory
P.O. Box 500
Mail Stop 105
Batavia, IL 60510

Paul C. Messina
Argonne National Laboratory
Building 221
9700 South Cass Avenue
Argonne, IL 60439

APPENDIX C
Bibliography

1. "Federal Plan for High-End Computing: Report of the High-End Computing Revitalization Task Force (HECRTF)," (Arlington, VA: National Coordination Office for Information Technology Research and Development, May 10, 2004), http://www.house.gov/science/hearings/full04/may13/hecrtf.pdf.
2. Horst Simon et al., "National Facility for Advanced Computational Science: A Sustainable Path to Scientific Discovery." Lawrence Berkeley National Laboratory report LBID-2509, April 8, 2004. A public version of the NFACS proposal without proprietary details, LBNL/PUB-5500, is available at http://www-library.lbl.gov/docs/PUB/5500/PDF/PUB-5500.pdf.
3. C. William McCurdy et al., "Creating Science-Driven Computer Architecture: A New Path to Scientific Leadership," Lawrence Berkeley National Laboratory report LBNL/PUB-5483, October 2002; http://www.nersc.gov/news/blueplanet.html.
4. "Facts on ASCI Purple," Lawrence Livermore National Laboratory report UCRL-TB-150327 (2002); http://www.sandia.gov/supercomp/sc2002/flyers/SC02ASCIPurplev4.pdf.
5. Daniel A. Reed, ed., "Workshop on the Roadmap for the Revitalization of High-End Computing," June 16–18, 2003 (Washington, D.C.: Computing Research Association).
6. Phillip Colella, Thom H. Dunning, Jr., William D. Gropp, and David E. Keyes, eds., "A Science-Based Case for Large-Scale Simulation" (Washington, D.C.: DOE Office of Science, July 30, 2003).
7. "Red Storm System Raises Bar on Supercomputer Scalability" (Seattle: Cray Inc., 2003), http://www.cray.com/company/RedStorm_flyer.pdf.

A1. David H. Bailey et al., "The NAS Parallel Benchmarks." Intl. Journal of Supercomputer Applications, vol. 5, no. 3 (Fall 1991), pg. 66–73.

A2. P. A. Agarwal et al., "Cray X1 Evaluation Status Report," ORNL Technical Report ORNL/TM-2004/13, January 2004.

A3. L. Oliker et al., "A Performance Evaluation of the Cray X1 for Scientific Applications," High Performance Computing for Computational Science VECPAR 2004, to appear.

APPENDIX D
Acronyms and Abbreviations

AMR Adaptive mesh refinement (numerical technique)
APDEC Applied Partial Differential Equations Center (SciDAC project)
ASCI Advanced Simulation and Computing (DOE/NNSA program)
BGK Bhatnagar, Gross and Krook (fusion)
BGL Blue Gene/L (IBM-LLNL research system)
CCSM Community Climate System Model (climate modeling)
CENIC Corporation for Education Network Initiatives in California
CITRIS Center for Information Technology Research in Interest of Society
CPU Central processing unit
DARPA Defense Advanced Research Projects Agency
DCA Dynamical cluster approximation (computational chemistry)
DOE U.S. Department of Energy
DSD Distributed Systems Department
EECS Department of Electrical Engineering and Computer Science (U.C. Berkeley)
EMT Extended Memory Technology (Intel)
ERCAP Energy Research Computing Allocations Process
ES Earth Simulator
ESnet Energy Sciences Network
ESSL Engineering and Scientific Software Library (IBM product)
Flop/s Floating-point operations per second
FFT Fast Fourier transform
FFTW (self-tuning FFT software)
FTE Full-time equivalent
FY Fiscal year
GB Gigabyte
GB/s Gigabytes per second
Gflop/s Gigaflop/s (billion floating-point operations per second)
GHz Gigahertz
GPFS General Parallel File System (IBM product)
GTC Stellarator Monte Carlo Transport (fusion code)
GYRO (gyrokinetic fusion code)
HCA Hardware custom accelerators (IBM design)
HECRTF High-End Computing Revitalization Task Force (multi-agency working group)
HOPI Hybrid Optical/Packet Infrastructure (Internet2 Working Group)
HPC High performance computing
HPCRD High Performance Computing Research Department (Berkeley Lab)
HPCS High Productivity Computing Systems
HPSS High Performance Storage System (IBM storage system)
IB InfiniBand (network technology)
INCITE Innovative and Novel Computational Impact on Theory and Experiment
I/O Input/output
ITER International Thermonuclear Experimental Reactor (fusion program)
JAMSTEC Japanese Marine Science and Technology Center
LAN Local area network
LBNL Lawrence Berkeley National Laboratory
LCRM Livermore Computing Resource Management (LLNL software)
LLNL Lawrence Livermore National Laboratory (DOE laboratory)
LSMS Locally Self-consistent Multiple Scattering (computational chemistry)
LU Lower-upper diagonal (numerical linear algebra technique)
MAN Metropolitan Area Network
MASS Mathematical Acceleration SubSystem (a math and science library from IBM)
MB Megabyte
MICS Mathematics, Information and Computer Science (DOE program)
MPEG Moving Pictures Experts Group (data compression standard)
MPI Message Passing Interface (parallel computing software)
MPLS Multiprotocol Label Switching (network technology)
NAS Numerical Aerospace Simulation (NASA Ames computer facility)
NASA National Aeronautics and Space Administration
NCSA National Center for Supercomputer Applications (NSF facility)
NERSC National Energy Research Scientific Computing Center
NIM NERSC Information Management (account and allocation software)
NIMROD Non-Ideal Magnetohydrodynamics with Rotation, Open Discussion (fusion)

NNSA National Nuclear Security Administration (DOE program)
NPB NAS Parallel Benchmarks
NREN NASA Research and Education Network (network)
NSF National Science Foundation
NWCHEM Northwest Chemistry (PNNL software)
OASCR Office of Advanced Scientific Computing Research (DOE program)
OC Optical cable (networking standard)
ORNL Oak Ridge National Laboratory (DOE laboratory)
OSF Oakland Scientific Facility (LBNL computer center)
PACI Partnership for Advanced Computational Infrastructure (NSF program)
PB Petabyte
PDE Partial differential equation (numerical approach)
PERC Performance Evaluation Research Center (SciDAC program)
PERCS Productive, Easy-to-use, Reliable Computing System (IBM project)
Pflop/s Petaflop/s (quadrillion floating-point operations per second)
PNNL Pacific Northwest National Laboratory (DOE laboratory)
POP Parallel Ocean Program (climate modeling code)
QMC Quantum Monte Carlo
QoS Quality of service (network technology)
RDMA Remote direct memory access
RFI Request for Information
SAN Storage area network
SC DOE Office of Science
SCaLeS Science Case for Large-scale Simulation (DOE working group)
SciDAC Scientific Discovery through Advanced Computing
SDSC San Diego Supercomputer Center (NSF facility)
SIMD Single instruction/multiple data
SLURM Simple Linux Utility for Resource Management (LLNL software)
SMP Symmetric multiprocessor
SP Scalable Parallel (IBM parallel computer product)
SSE Streaming SIMD Extensions (Intel)
SuperLU (numerical software product for sparse LU factorization)
TB Terabyte
TeraGrid (NSF distributed facility)
Tflop/s Teraflop/s (trillion floating-point operations per second)
TLBE Thermal Lattice Boltzmann Equation (fusion code)
UC University of California
UPC Unified Parallel C (programming language)
ViVA Virtual Vector Architecture (LBNL-IBM project)
VLIW Very long instruction word (computer architecture)
WAN Wide area network
WRF Weather Research and Forecasting Model